Holonic Science: A New Framework for Benchmarking AI Scientists

Nathan Suri
Department of Physics
Yale University
New Haven, CT 06511
nathan.suri@yale.edu

Savannah Thais
Department of Computer Science
Hunter College
New York, NY 10065
savannah.thais@hunter.cuny.edu

Abstract

Science is a system defined in part by measurability. Claims made under its banner are trusted under the implicit understanding that they can be verified through measurement. Trustworthy science is therefore only possible when accurate and verifiable measurements of all aspects of a discovery or observation are possible. Recently, a new interloper has emerged in the form of AI scientists. Driven by companies such as Sakana AI and Google, these hybrid human-AI systems tasked with scientific discovery strive to augment and accelerate the current research paradigm by intelligently innovating upon and combining preexisting ideas. As researchers attempt to build collaborative workflows with AI scientists, the need for better measurements of their capabilities and limitations escalates. In this paper, we argue that the complexity of scientific research represents a significant challenge to AI scientist benchmarking attempts on account of construct validity issues. Scientific research tasks must be parseable by AI scientists, otherwise these in silico collaborators pose a significant epistemic risk to the trustworthiness of scientific research. To address this, we propose a new framework for designing benchmarks for AI scientists based on Arthur Koestler's concept of holons. Instead of benchmarking high-level human-interpretable tasks, we instead break them down and build specialized benchmarks at the LLM-executable level. The semantic sum of an AI scientist's performance on these benchmarks will then approximate performance on the original task. Our framework outlines key criteria for future benchmarks to avoid construct validity issues. We also exemplify the potential of our framework by prototyping a benchmark for attributional accuracy ultimately aimed at evaluating AI scientists on their ability to generate literature reviews.

1 Introduction

Science is a human-designed system that exists to engage with and explain the natural universe defined by our ability to measure. The scientific community has evolved into a knowledge ecosystem joined by the mission that any claim can be validated due to its inherent measurability. Science is therefore only "scientific" when accurate and verifiable measurements are possible; for it is the measurability of scientific results that differentiates modern science from its medieval antecedents like alchemy and astrology. The critical subtlety here is that trustworthy science is only possible if *all* steps of the scientific process are verifiably measurable and documentable. However, this imperative now finds itself challenged by rising interest in utilizing hybrid human-AI systems known colloquially as 'AI scientists' to automate scientific discovery workflows.

AI systems have already proven to be useful inclusions in the pursuit of scientific discoveries. Two of the most recent Nobel Prizes (Chemistry, 2024 [14] and Physics, 2025 [13]) have exemplified the value provided by AI models in performing complex computational scientific tasks. As such,

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Science.

it is no surprise that scientists have pursued even greater degrees of collaboration with AI systems, ultimately seeking to fully automate the whole scientific discovery workflow. Numerous works across multiple fields of study have sought to demonstrate that large language models (LLMs) are capable of performing scientific discovery tasks such as scientific contextualization [6], problem specification [33][27][36], hypothesis generation [2], experimental design [3][5], and evaluation [34][16]. However, these attempts are still a far cry from proving that LLMs can perform trustworthy science. Notably, these prototypal automated workflows struggle with specialized scientific reasoning, long-term iterative planning, and critical analysis in collaborative workflows [21]. Fully integrating AI systems in scientific discovery workflows can then result in critical epistemic risks for scientific research. Unlike in test scenarios, real data is often incomplete or imprecise, which can result in emergent errors in data analysis and conclusion drawing [7]. Giving AI systems access to physical experimental setups can lead to hazardous consequences resulting from departures from intended safety measures [30]. Generally speaking, outsourcing more agency to these systems allows for the possibility of biasing research output, propagating unreliable results, and experimental system failures [10]. To ensure that the automation of high-level human-interpretable scientific discovery tasks by AI scientists will not result in emergent risks to scientific integrity, scientists must be able to clearly define each task in a measurable manner that is both representative of the original and executable by an LLM.

The challenge arises from the fact that scientific discovery at a high level is not easily discretized. More explicitly, high-level human-interpretable tasks within scientific discovery workflows such as literature review or experimental design are poorly mapped to the space of low-level LLM-executable tasks such as searching or fact checking. Thus, any misalignment between the AI scientists' output and the desired outcome stems from the former's inability to fully conceptualize and comprehend the intent behind any step in the scientific discovery cycle, fundamentally a construct validity issue [20]. As a rational call for better guardrailing, we posit that the danger of AI scientists comes not from simply using them, but using them without sufficient comprehension of our expectations for a task. No tool is ever ideal, but by measuring its deviation from the optimal outcome, we can bound our expectations through quantifiable uncertainties.

Benchmarks serve as the primary method by which researchers can study the performance and limitations of AI systems on complex tasks. These evaluation workflows typically consist of a cultivated dataset made up of a taxonomized list of tasks and a series of metrics to test the aptitude of AI systems [20]. As the concept of collaborating with AI scientists transitions from being prototypal to commonplace, it becomes increasingly vital to design realistic benchmarks to measure the capacity that AI scientists have to perform common tasks in the scientific discovery workflow.

However, these tasks are not easily disentangled into discrete, regulatable actions executable by LLMs. Rather, scientific discovery is a dynamic workflow of indeterminate human-interpretable steps tied together via highly complex reasoning and perception that constantly evolves as new technologies are introduced. Many existing efforts to benchmark AI scientists side-step this difference by evaluating human-interpretable tasks by constructing and measuring performance on limited and potentially misleading reductions that are unrealistic and unrepresentative of the original task. For example, in attempting to assess citation accuracy in the context of literature reviews, the CiteME benchmark is comprised of queries to identify a referenced paper given a text excerpt with a masked in-text citation [19]. While an improvement over earlier retrieval-based benchmarks [11], this task construction is not representative of realistic use cases of AI systems as research assistants and thus indicative of a construct validity issue. It should be noted that concerns over construct validity are not exclusive to AI scientist benchmarks. Other domains involving complex tasks such as legal reasoning face similar misalignment obstacles between claims of general task mastery and the representativeness of the benchmark's task set [12].

To address the construct validity issue for AI scientist benchmarks, we propose a framework for breaking down aspects of scientific discovery into valid benchmark tasks derived from Arthur Koestler's descriptions of holons in *The Ghost in the Machine* [15]. Our framework underscores that performance on human-interpretable tasks is approximately equivalent to the semantic sum of assessments on LLM-executable sub-tasks or holons. Each holon defines its own problem space to be evaluated, but remains a part of a larger system to measure more general reasoning capabilities. By constructing holonic benchmarks that are representative of specific constructs within scientific research tasks, yet are valid evaluations in their own rights, we can strive for collaborative settings in which AI scientists will interpret any high-level task in alignment with our expectations for scientific research. Section 4 exemplifies the power of our framework by using it to sketch a benchmark for attributional accuracy and show how it can be used to evaluate the AI scientist's upstream ability to generate realistic and useful literature reviews.

2 AI Scientists

As this paper seeks to define a new framework for benchmarking AI scientists, it is imperative that we explicitly define what types of hybrid human-AI systems we seek to evaluate. Conventionally, the term AI scientists refers to workflows involving AI agents that autonomously perform scientific research tasks. However, as the concept is nascent, this description remains a convention rather than any strict denotation. AI systems have been used as in silico collaborators to perform a wide range of tasks from code review to hypothesis generation [21][9]. Despite their differences in scope, each setup has the potential to obfuscate the valid measurability of the scientific discovery workflow. To encompass all possible collaborative setups, we opt for the broadest possible definition of an AI scientist, while also accounting for the spectrum of possible human steering present:

Definition 2.1. An AI scientist refers to any hybrid human-AI workflow that relies on AI for scientific discovery assistance beyond simple quantitative analyses. We further subdivide the category into three mutually exclusive types of AI scientists based on the degree of human steering present.

1. Human-AI assistant: The research workflow remains in its traditional, human-driven form augmented with aid from AI assistants such as chatbots.
2. Human-AI agent: Specific steps of the research process are outsourced to singular agents with a human overseer. When multiple, decoupled agents are used, the burden of aggregating and adjudicating the results lies with the human overseer. Here, the singular agents are equivalent to what are conventionally known as AI scientists: autoregressive large language models guided by agentic frameworks to adopt reasoning akin to those exhibited by human researchers [9].
3. Multi-agentic system: (Nearly) all steps of the research process are outsourced to agents with a central agent overseer and a human adjudicator [9][35].

Across multiple fields of study, scientists have already shown the potential benefits of the ""scientist-in-the-loop" collaborative paradigm" [9]. Coscientist, a GPT-4 based system for autonomous chemical research, was able to successfully optimize the reaction of palladium-catalysed cross-couplings [5]. CellAgent, another GPT-4 agent, has been shown to automate single-cell RNA sequencing data analysis more reliably than its contemporaries [34]. PriM is a multi-agent system designed for automated materials discovery that yields high materials exploration rate without significant reduction in scientific rigor [16].

In spite of the benefits, integrating AI into scientific research pipelines proves to be a double-edged sword replete with significant epistemic risks. It has already been demonstrated that AI agents struggle with computational reproducibility, a critical method for maintaining the integrity of scientific research [28]. Additionally, the fact that many AI scientists are built using privately trained large language models (LLMs) makes it difficult to determine any possible biases in their outputs and whether they are inherited from their training corpora [32] or programmed preference optimization schema [18]. The problem is not limited to agentic AI, as baseline LLMs continue to struggle on benchmarks of latent constructs critical to scientific discovery such as mathematical reasoning [25]. Beyond experimentally verified risks, AI researchers also forecast emergent risks such as over-reliance among human users [18] and self-preservation in conflict with human interests [4].

3 Benchmarking Scientific Research

While scientists can rigorously answer questions about the quantifiable validity of their discoveries, explicitly defining the required research steps to produce these results proves to be more elusive of a goal. Due to the complicated reality of scientific discovery, taxonomies of scientific research tasks such as the "scientific method" are typically left as instructional tools and thus vary greatly in diction [8]. However, if we as scientists cannot establish a reasonable list of scientific research tasks and their respective aims, expecting AI scientists to perform accurately on these high-level human-interpretable tasks is overstepping the mark.

Scientific research tasks pose significant challenges when serving as the target for AI benchmarks. Having been in use since the 1970s [20], AI benchmarks have acquired a conventional construction consisting of a training dataset, queries, and a series of evaluation metrics to assess an AI model on certain tasks. Here, tasks refer to any mapping between a problem and action space with intensional (human-interpretable task description) and extensional (empirical pairs of states and actions) definitions [26]. However, unlike strictly computational tasks with defined problem-action mappings such as image classification [23], scientific research tasks are ill-defined in scope, frequently with problem spaces mapping to other problem spaces before any notion of actions are relevant. Under such unclear construction, these high-level hierarchical tasks introduce significant variability among the possible trajectories from original problem space to outcome states on account of the actor's interpretation of the task [26]. While humans can sort through the interconnected and multifaceted dependencies of scientific research tasks intuitively, it is currently unknown if or to what degree AI scientists could. While highlighted here for AI scientist benchmarks, this mismatch between intensional and extensional definitions manifests as construct validity issues in any assessment of high-level linguistic competence [20] including but not limited to legal reasoning [12], diagnosing mental disorders [29], and financial support [31]. Without clear and specific definitions of outcome expectations, AI systems are prone to misalignment with respect to our desires as collaborators.

For the sake of exemplifying our proposed framework, we define the following taxonomy of research tasks required for scientific discovery. Due to the ad hoc nature of scientific discovery, a perfect enumeration of tasks is not achievable. As we will argue later on, the construction of a taxonomy of an overall concept such as scientific discovery is subjective and thus requires justification. While other task breakdowns are theoretically possible, we chose this taxonomy due to its simplicity due to defining relatively exclusive tasks within the space of scientific discovery. It should be noted that in practice, scientific discovery tasks are often done in parallel with many interwoven dependencies, thus requiring domain expert input to integrate this relationships into a more involved taxonomy.

- Literature review: Detailed and extensive knowledge retrieval and contextualization within the current status quo of the field of study
- Problem specification: Identification of a novel knowledge gap or an unexplained observation to be formulated into a research project
- Research design: Theorizing of a potential solution to the selected research problem that is both empirically testable and potentially falsifiable.
- Experimentation: Encompasses all aspects of experimental design and execution, whether it be computational or physical in nature
- Communication: All steps required for the dissemination of scientific results via the peer review publication process

3.1 Holonic Solution

A critical limitation of benchmarks aimed at assessing the capability of AI systems to perform language tasks is the conceptual gap between the evaluation framework and the intended human-interpretable task to be completed [24]. The scientific discovery workflow is no exception. Due to their inherent complexity, any task enumerated in Section 3 necessitates the model to exhibit general scientific reasoning. Each human-interpretable task corresponds to a collection of trajectories between sub-tasks and possible outcomes with no one precise solution. This variability poses a fundamental dilemma to the construction of realistic benchmark tasks in closing the aforementioned conceptual gap. Conventional benchmarks opt for reducing human-interpretable tasks to simpler versions with easier (albeit less representative) datasets to compile [24]. However, these reduced tasks frequently end up being contrived "samples of convenience: tasks and collections of tasks arbitrarily built out of what is easily available to the team developing these benchmarks, even if such constructions are theoretically unsound" [20]. For example, instead of directly tackling the corresponding human-interpretable task of literature review according to our taxonomy in Section 3, the CiteME benchmark narrows in on evaluating an LLM's capacity for citation attribution. CiteME chooses to model queries of citation attribution as a retrieval task for a referenced paper from a text excerpt with a masked in-text citation [19]. While the intensional and extensional definitions are consistent, their construction is not representative of real attributional queries by human researchers, who usually are not validating known citations in text excerpts. Without attempting to realistically

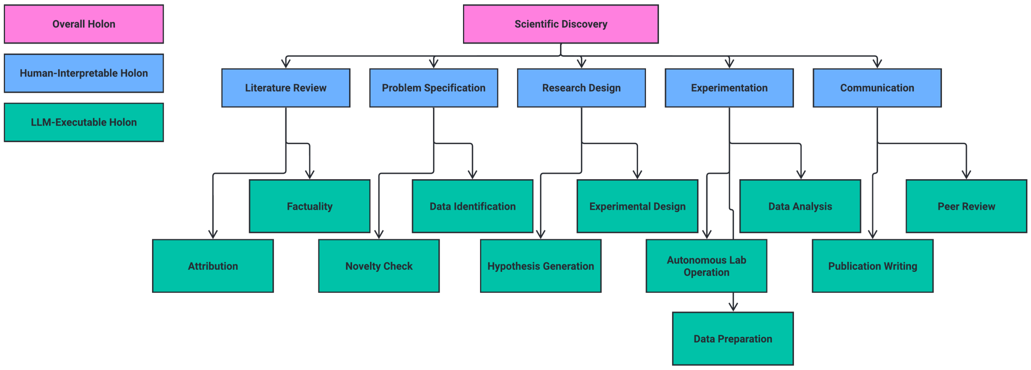

Figure 1: Prototypal holonic breakdown of the scientific discovery workflow

reduce the semantic distance between real human-interpretable tasks and the simulated tasks in the benchmark, the conventional development framework fails to accurately specify the expected trajectory and outcome of the human-interpretable task and by extension, fails to accurately evaluate the model on its aptitude for scientific research.

To address this construct validity issue, we propose a new framework for developing benchmarks for AI scientists based on a holonic perspective on scientific research. In *The Ghost in the Machine*, Arthur Koestler defined a holon as an identifiable subsystem with a unique identity such that it is both a self-contained system in its own right as well as part of a larger holarchical structure [15]. We posit that scientific research as a whole can be interpreted as a holarchical structure with the largest holon being the task of scientific discovery, further divided into intermediate holonic human-interpretable tasks like those enumerated in Section 3. While this breakdown is sufficient for human researchers, we argue that another layer of splitting is required for AI scientists. This layer further subdivides the human-interpretable holons into holonic tasks that are executable by LLMs. Rather than constructing a single benchmark for each human-interpretable task, our framework poses that benchmarks should be built at the LLM-executable level. By only building benchmarks for LLM-executable tasks, we minimize the conceptual gap between the evaluation framework (extensional definition) and the intensional definition of the task. Under this methodology, we rely on the combined evaluations of each LLM-executable benchmark to approximate the AI scientist's aptitude for any (human-interpretable) scientific research task. While the constituent holons themselves are still narrowly defined, by creating summative evaluation frameworks at the human-interpretable level, we become more adept at validating trajectories of AI scientists between problem spaces and subspaces for the complex tasks described in Section 3. Figure 3.1 illustrates how our framework maps the taxonomy of scientific discovery tasks chosen in Section 3 to a holonic breakdown all the way to the LLM-executable level. Section 4 will exemplify the theoretical utility of our framework by outlining a benchmark design to assess the attributional aptitude of AI scientists.

As benchmarks in their own right, the LLM-executable holons are subject to the same validity issues as contemporary benchmarks. In Koestler's words, regardless of the holarchical structure, "constituent holons are defined by fixed rules and flexible strategies" [15]. As such, we require each LLM-executable benchmark to exhibit the following criteria inspired by the forms of validity discussed in Salaudeen et al. [24]:

1. Focused: The central task(s) of the benchmark should be highly specialized in scope to maintain high reproducibility and reliability.
2. Realistic: The task(s) should be constructed in accordance with actual use cases sourced from human researchers.
3. Measurable in depth: The benchmark should evaluate its task(s) with a variety of metrics, not only for technical accuracy but also for alignment and robustness against adversarial use cases.
4. Scale independence: The task(s) should be measurable across all AI scientist workflows regardless of the degree of human steering present.

5. Domain generalizability: The benchmark dataset should, if relevant, contain queries from multiple scientific domains with domain-specific benchmarks requiring stronger justification.

4 Attribution

With our framework defined, we can now showcase how this could apply in practice by prototyping a benchmark for a single LLM-executable holon: attribution. Ultimately, this benchmark could be in conjunction with other LLM-executable benchmarks to assess the ability of an AI scientist to generate realistic and useful literature reviews.

Accurate attribution is a requirement for properly situating one's work and potential future ideas within the existing knowledge ecosystem of scientific knowledge. Beyond communicating directly with the primary researcher(s), attribution provides the only public avenue for tracing the provenance of any scientific result via the listed citations in a peer-reviewed publication. If these citations are unethically compiled, peer scientists lose a critical means by which to assess the results of another party as well as a starting point for their own literature reviews. Attribution however is not limited to assessments of content mapping accuracy, but can extend to more qualitative requirements. For example, in performing a literature review, scientists need to acquire information from sources that are modern, unbiased, and diverse in methodology [22]. Thus, an AI scientist would also need to recommend a variety of sources that allow for a comprehensive survey of a topic landscape. As such, we extend the definition of attributional accuracy to encompass both evaluations of accuracy and unbiased utility.

While the importance of literature reviews is commonly overshadowed by experimentation or evaluation, it remains a vital foundation for scientific discovery. Within the system of science, literature reviews serve as the primary mechanism by which the scientific community can referentially engage with each other by way of the compendium of scientific knowledge. For researchers, the aim of literature reviews is to survey relevant literature to engender original ideas and logically support actions taken in the pursuit of scientific discovery. Here, the underlying principle of measurability extends not only to empirical observations of the physical world, but also the congruence of one's work with the existing literature. As the desire to fully automate the scientific discovery workflow escalates, we must advocate for the construction of realistic benchmarks for all constituent human-interpretable tasks including literature review. For without ensuring accurate literature reviews, how can we hope to protect the integrity of science?

Literature reviews seek to survey and analyze scientific literature in order to "synthesise current knowledge, critically discuss existing proposals, and identify trends" [6]. Accurate literature reviews should be well-defined in scope, justified in its inclusion and exclusion criteria, comprehensive, and unbiased [22]. Each of these criteria further complicate a task already challenging for human researchers, let alone for AI scientists. We argue that the problem space defined by the act of generating accurate and useful literature reviews proves to be too large to easily map onto action spaces in one step. This inevitably leads to a conceptual gap between the intended human-interpretable task of literature review and any attempt at designing a representative benchmark, therefore denoting a construct validity issue.

To exemplify the validity issues present with the conventional approach to benchmarking AI scientists, we review contemporary benchmarks that seek to evaluate similarly defined tasks. CiteME is a citation-based benchmark that focuses on asking whether LLMs are capable of correctly identifying a target paper from masked in-text citation in a text excerpt from a source paper [19]. However, the citation attribution task designed for CiteME is a limited construction. The task is an ex post facto verification of attribution, essentially side-stepping the full task of citation attribution in favor of semantic searching. As such, we find the task definition in the CiteME benchmark as not realistic. LitSearch is a retrieval benchmark that primarily seeks to assess non-LLM retrieval systems on citation recommendation queries [1]. The task construction of LitSearch differs from CiteME by sampling inline target citations from source papers and then using GPT-4 to generate natural language information requests to guide its citation recommendations. The authors of LitSearch also manually reviewed the queries to ensure they were rigorous by removing any that were too close semantically to the target paper title. However, LitSearch was aimed at non-LLM retrieval systems, reducing its applicability in evaluating a broad spectrum of AI scientists. Additionally, the datasets of both CiteME and LitSearch only contain queries pertaining to ML and NLP papers specifically.

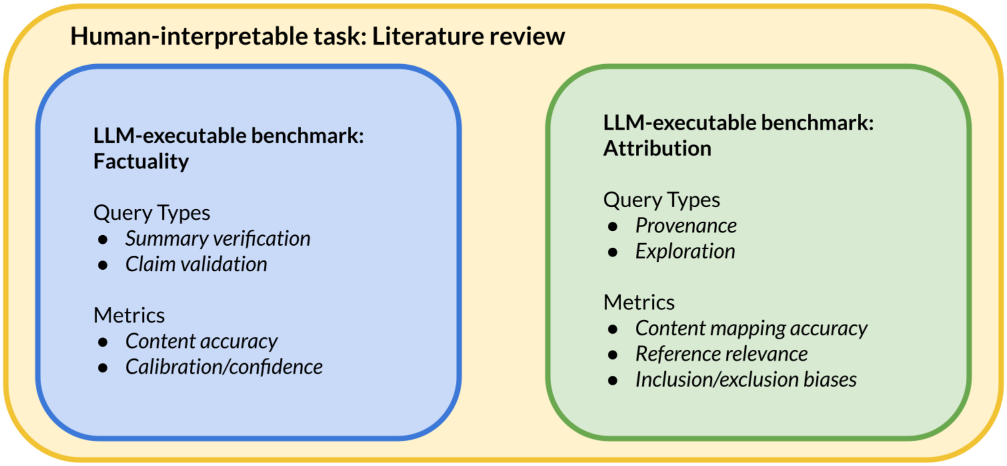

Figure 2: Proposed holonic structure for benchmarking literature review

While CiteME and LitSearch struggled to overcome the validity issues posed by benchmarking the human-interpretable task of literature review, it is through such complex task setups that our proposed holonic framework for benchmark development shines. Using the conventions laid out in Section 3.1, instead of directly constructing a benchmark for the human-interpretable holon of literature review, we divide the latter into the LLM-executable holons of factuality and attribution. Figure 4 illustrates how our framework can be used to more effectively evaluate the human-interpretable task of literature review as the semantic sum of LLM-executable holonic benchmarks. Here, factuality refers to queries of content accuracy. A factuality benchmark would focus on assessing whether summaries of scientific works present content that is accurate to the modern knowledge ecosystem of a scientific discipline. Attribution refers to queries of content mapping aptitude. An attributional benchmark would seek to evaluate both the relevance of provided source materials as well as the accuracy of the mapping between intellectual content to its original source.

4.1 Benchmark Design

Having broken down the human-interpretable task of literature review to the LLM-executable level, the tasks at the holonic level prove to be specialized enough to ensure that each benchmark is focused. For the attributional benchmark, each query will be constructed as a request for a set of sources relevant to provided information in the prompt. The dataset will then consist of pairings of prompts defining the request space and appropriate listings of relevant sources defining the output space. In order for our queries to be realistic, we will request feedback from human scientists who either currently collaborate with or are interested in collaborating with AI scientists. This feedback will guide our construction of the queries that will serve as the underlying dataset by aligning the diction and syntax of these queries with real or desired use cases. Additionally, we will require that the queries sample sources from a variety of scientific domains.

We identify two distinct types of queries based on the size of the desired output space: provenance and exploration. The task of provenance is defined as a mapping between a request space populated by information requests and an output space of low dimensionality (O($k \approx 1$)) populated by potential references. The query will be framed as either directly asking for a specific source or requesting specific information that maps to a low number of potential sources. The output of provenance tasks will be structured as a ranked list of k sources, ordered based on their relevance to their initial query. Similarly to provenance queries, the task of exploration is defined as a mapping between a request space populated by information requests and an output space of unbounded dimensionality populated by relevant references. The final iteration of the task will also be a ranked list of k sources, where a reasonable k will be chosen to represent typical reference counts requested by human researchers.

Despite the focused task set, we can maximize the utility of the benchmark by assessing each task with a variety of metrics. Specifically, we will choose metrics to determine the likelihood and severity by which an AI scientist could exhibit attributional malpractices. We define three risk areas for

Table 1: Taxonomized list of attributional malpractices

Risk Areas		**Malpractices**
Mistake		Fabrication of content (plagiarism) Fabrication of sources Mis-mapping between sources
Misuse		Injected biases Direct output sculpting
Misalignment	Misinterpretation	Source incomprehension Source misrepresentation
	Biased Attribution	Intrinsic biases Learned strategies

attributional malpractices: mistakes (technical failures), misuse (adversarial exploitation by human collaborators), and misalignment (knowingly acting against the intent of human collaborators). Table 1 presents a taxonomy of attributional malpractices segregated between the three risk areas. These malpractices could be measured by assessing the impact of propagated biases, model robustness to prompt perturbations, and through adversarial simulations [17]. We will also test our benchmark across the multiple scales of AI scientist workflows defined in Section 2 to quantify their potential risk factors.

5 Conclusions

AI scientists have the potential to completely revolutionize the pursuit of scientific truths forever. With explosive rates of publications, the rise of the big data paradigm, and a growing desire for automation in general, it is no surprise that scientists across multiple fields of study have sought to harness the power of LLMs to augment their capabilities for scientific discovery. As such, it is a futile effort to obstruct the oncoming wave of AI scientist efforts. We have already seen significant preliminary benefits from human-AI collaborations [5][34][16]. However, the immense power of AI scientists should beget an even greater responsibility among current scientists to protect the integrity of scientific research.

In order to determine and prevent the potential risks posed by AI scientists to the scientific discovery workflow, we must act scientifically by measuring the aptitude of AI scientists on scientific research tasks. Unfortunately, scientific research is an ever-evolving series of ill-defined tasks that prove challenging even for human researchers to navigate. The conceptual gap posed by improper communication of the expectations of a task's outcome to an AI scientist can lead to critical failures in workflows it was intended to accelerate. Attempting to benchmark tasks at this high level of abstraction results in construct validity issues [20]. Contemporary benchmarks fail to overcome this challenge, opting to generate task sets reduced in scope that are more manageable and thus more easily measurable. However, these reduced tasks frequently are not representative of their intended real-world equivalent. Our paper proposes a new framework for developing benchmarks based on Arthur Koestler's holons [15]. Instead of building benchmarks for human-interpretable tasks such as literature review, we opt to construct benchmarks for LLM-executable holons, whose semantic sum becomes approximately equivalent to the original human-interpretable task. Each LLM-executable holon is robust enough to warrant its own benchmark, but specific enough to better ensure the reproducibility and validity of the results.

To exemplify the power of our framework, we sketched a prototypal benchmark for attribution, an LLM-executable holon for the human-interpretable task of literature review. When completed, the benchmark promises to be the first benchmark to assess the attributional accuracy of AI scientists at all scales of human steering with cross-disciplinary queries. In conjunction with this paper, we hope that the benchmark will inspire further efforts to benchmark AI scientists as they become ubiquitous additions to the scientific discovery workflow.

References

[1] Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. LitSearch: A Retrieval Benchmark for Scientific Literature Search, October 2024. URL `http://arxiv.org/abs/2407.18940`. arXiv:2407.18940 [cs].

[2] Atilla Kaan Alkan, Shashwat Sourav, Maja Jablonska, Simone Astarita, Rishabh Chakrabarty, Nikhil Garuda, Pranav Khetarpal, Maciej Pióro, Dimitrios Tanoglidis, Kartheik G. Iyer, Mugdha S. Polimera, Michael J. Smith, Tirthankar Ghosal, Marc Huertas-Company, Sandor Kruk, Kevin Schawinski, and Ioana Ciucă. A Survey on Hypothesis Generation for Scientific Discovery in the Era of Large Language Models, April 2025. URL `http://arxiv.org/abs/2504.05496`. arXiv:2504.05496 [cs] version: 1.

[3] Sören Arlt, Haonan Duan, Felix Li, Sang Michael Xie, Yuhuai Wu, and Mario Krenn. Meta-Designing Quantum Experiments with Language Models, July 2025. URL `http://arxiv.org/abs/2406.02470`. arXiv:2406.02470 [quant-ph].

[4] Yoshua Bengio, Michael Cohen, Damiano Fornasiere, Joumana Ghosn, Pietro Greiner, Matt MacDermott, Sören Mindermann, Adam Oberman, Jesse Richardson, Oliver Richardson, Marc-Antoine Rondeau, Pierre-Luc St-Charles, and David Williams-King. Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?, February 2025. URL `http://arxiv.org/abs/2502.15657`. arXiv:2502.15657 [cs].

[5] Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, December 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06792-0. URL `https://www.nature.com/articles/s41586-023-06792-0`. Publisher: Nature Publishing Group.

[6] José de la Torre-López, Aurora Ramírez, and José Raúl Romero. Artificial intelligence to automate the systematic review of scientific literature. *Computing*, 105(10):2171–2194, October 2023. ISSN 1436-5057. doi: 10.1007/s00607-023-01181-x. URL `https://doi.org/10.1007/s00607-023-01181-x`.

[7] Maria Eriksson, Erasmo Purificato, Arman Noroozian, Joao Vinagre, Guillaume Chaslot, Emilia Gomez, and David Fernandez-Llorca. Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation, February 2025. URL `https://arxiv.org/abs/2502.06559v2`.

[8] Daniel Trugillo Martins Fontes. Nature of Science in the classroom: empirical insights into investigating physics. *A Física na Escola*, 22:230136, August 2024. ISSN 1983-6430, 1983-6422. doi: 10.59727/fne.v22i1.136. URL `http://arxiv.org/abs/2504.03912`. arXiv:2504.03912 [physics].

[9] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R. D. Costa, José R. Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Towards an AI co-scientist, February 2025. URL `http://arxiv.org/abs/2502.18864`. arXiv:2502.18864 [cs].

[10] Mourad Gridach, Jay Nanavati, Khaldoun Zine El Abidine, Lenon Mendes, and Christina Mack. Agentic AI for Scientific Discovery: A Survey of Progress, Challenges, and Future Directions, March 2025. URL `http://arxiv.org/abs/2503.08979`. arXiv:2503.08979 [cs].

[11] Nianlong Gu, Yingqiang Gao, and Richard H. R. Hahnloser. Local Citation Recommendation with Hierarchical-Attention Text Encoder and SciBERT-based Reranking, March 2022. URL `http://arxiv.org/abs/2112.01206`. arXiv:2112.01206 [cs].

[12] Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson,

Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models, August 2023. URL `http://arxiv.org/abs/2308.11462`. arXiv:2308.11462 [cs].

[13] J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, April 1982. doi: 10.1073/pnas.79.8.2554. URL `https://www.pnas.org/doi/10.1073/pnas.79.8.2554`. Publisher: Proceedings of the National Academy of Sciences.

[14] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL `https://www.nature.com/articles/s41586-021-03819-2`. Publisher: Nature Publishing Group.

[15] A. Koestler. *The Ghost in the Machine*. An Arkana book : philosophy. Penguin Group (USA) Incorporated, 1989. ISBN 978-0-14-019192-9. URL `https://books.google.com/books?id=tJS-QgAACAAJ`.

[16] Zheyuan Lai and Yingming Pu. PriM: Principle-Inspired Material Discovery through Multi-Agent Collaboration, April 2025. URL `http://arxiv.org/abs/2504.08810`. arXiv:2504.08810 [cs] version: 1.

[17] Lizhi Lin, Honglin Mu, Zenan Zhai, Minghan Wang, Yuxia Wang, Renxi Wang, Junjie Gao, Yixuan Zhang, Wanxiang Che, Timothy Baldwin, Xudong Han, and Haonan Li. Against The Achilles' Heel: A Survey on Red Teaming for Generative Models, November 2024. URL `http://arxiv.org/abs/2404.00629`. arXiv:2404.00629 [cs] version: 2.

[18] Margaret Mitchell, Avijit Ghosh, Alexandra Sasha Luccioni, and Giada Pistilli. Fully Autonomous AI Agents Should Not be Developed, February 2025. URL `http://arxiv.org/abs/2502.02649`. arXiv:2502.02649 [cs] version: 2.

[19] Ori Press, Andreas Hochlehnert, Ameya Prabhu, Vishaal Udandarao, Ofir Press, and Matthias Bethge. CiteME: Can Language Models Accurately Cite Scientific Claims?, November 2024. URL `http://arxiv.org/abs/2407.12861`. arXiv:2407.12861 [cs].

[20] Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the Everything in the Whole Wide World Benchmark, November 2021. URL `http://arxiv.org/abs/2111.15366`. arXiv:2111.15366 [cs].

[21] Chandan K. Reddy and Parshin Shojaee. Towards Scientific Discovery with Generative AI: Progress, Opportunities, and Challenges, December 2024. URL `http://arxiv.org/abs/2412.11427`. arXiv:2412.11427 [cs] version: 2.

[22] Meike Ressing, Maria Blettner, and Stefanie J. Klug. Systematic Literature Reviews and Meta-Analyses. *Deutsches Ärzteblatt International*, 106(27):456–463, July 2009. ISSN 1866-0452. doi: 10.3238/arztebl.2009.0456. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2719096/`.

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, January 2015. URL `http://arxiv.org/abs/1409.0575`. arXiv:1409.0575 [cs].

[24] Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. Measurement to Meaning: A Validity-Centered Framework for AI Evaluation, June 2025. URL http://arxiv.org/abs/2505.10573. arXiv:2505.10573 [cs] version: 4.

[25] Ankit Satpute, Noah Giessing, Andre Greiner-Petter, Moritz Schubotz, Olaf Teschke, Akiko Aizawa, and Bela Gipp. Can LLMs Master Math? Investigating Large Language Models on Math Stack Exchange, March 2024. URL http://arxiv.org/abs/2404.00344. arXiv:2404.00344 [cs] version: 1.

[26] David Schlangen. Targeting the Benchmark: On Methodology in Current Natural Language Processing Research. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.85. URL https://aclanthology.org/2021.acl-short.85/.

[27] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers, September 2024. URL http://arxiv.org/abs/2409.04109. arXiv:2409.04109 [cs].

[28] Zachary S. Siegel, Sayash Kapoor, Nitya Nagdir, Benedikt Stroebl, and Arvind Narayanan. CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark, September 2024. URL http://arxiv.org/abs/2409.11363. arXiv:2409.11363 [cs].

[29] Jie Sun, Tangsheng Lu, Xuexiao Shao, Ying Han, Yu Xia, Yongbo Zheng, Yongxiang Wang, Xinmin Li, Arun Ravindran, Lizhou Fan, Yin Fang, Xiujun Zhang, Nisha Ravindran, Yumei Wang, Xiaoxing Liu, and Lin Lu. Practical AI application in psychiatry: historical review and future directions. *Molecular Psychiatry*, 30(9):4399–4408, September 2025. ISSN 1476-5578. doi: 10.1038/s41380-025-03072-3. URL https://www.nature.com/articles/s41380-025-03072-3. Publisher: Nature Publishing Group.

[30] Xiangru Tang, Qiao Jin, Kunlun Zhu, Tongxin Yuan, Yichi Zhang, Wangchunshu Zhou, Meng Qu, Yilun Zhao, Jian Tang, Zhuosheng Zhang, Arman Cohan, Dov Greenbaum, Zhiyong Lu, and Mark Gerstein. Risks of AI scientists: prioritizing safeguarding over autonomy. *Nature Communications*, 16(1):8317, September 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-63913-1. URL https://www.nature.com/articles/s41467-025-63913-1. Publisher: Nature Publishing Group.

[31] Darko B. Vuković, Senanu Dekpo-Adza, and Stefana Matović. AI integration in financial services: a systematic review of trends and regulatory challenges. *Humanities and Social Sciences Communications*, 12(1):562, April 2025. ISSN 2662-9992. doi: 10.1057/s41599-025-04850-8. URL https://www.nature.com/articles/s41599-025-04850-8. Publisher: Palgrave.

[32] Oskar van der Wal, Dominik Bachmann, Alina Leidinger, Leendert van Maanen, Willem Zuidema, and Katrin Schulz. Undesirable Biases in NLP: Addressing Challenges of Measurement. *Journal of Artificial Intelligence Research*, 79:1–40, January 2024. ISSN 1076-9757. doi: 10.1613/jair.1.15195. URL http://arxiv.org/abs/2211.13709. arXiv:2211.13709 [cs].

[33] Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific Inspiration Machines Optimized for Novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 279–299, 2024. doi: 10.18653/v1/2024.acl-long.18. URL http://arxiv.org/abs/2305.14259. arXiv:2305.14259 [cs].

[34] Yihang Xiao, Jinyi Liu, Yan Zheng, Xiaohan Xie, Jianye Hao, Mingzhi Li, Ruitao Wang, Fei Ni, Yuxiao Li, Jintian Luo, Shaoqing Jiao, and Jiajie Peng. CellAgent: An LLM-driven Multi-Agent Framework for Automated Single-cell Data Analysis, July 2024. URL http://arxiv.org/abs/2407.09811. arXiv:2407.09811 [cs].

[35] Kevin G. Yager. Towards a science exocortex. *Digital Discovery*, 3(10):1933–1957, October 2024. ISSN 2635-098X. doi: 10.1039/D4DD00178H. URL https://pubs.rsc.org/en/content/articlelanding/2024/dd/d4dd00178h. Publisher: RSC.

[36] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search, April 2025. URL `http://arxiv.org/abs/2504.08066`. arXiv:2504.08066 [cs].